DOCUMENT RESUME

ED 400 332 TM 025 737

AUTHOR Bridgeman, Brent; And Others
TITLE Choice among Essay Topics: Impact on Performance and Validity.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY College Entrance Examination Board, New York, N.Y.
REPORT NO ETS-RR-96-4
PUB DATE Feb 96
NOTE 29p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Advanced Placement; *Essay Tests; European History; High Achievement; High Schools; *High School Students; *Performance Factors; *Scores; *Student Attitudes; Test Results; *Test Validity; United States History
IDENTIFIERS Advanced Placement Examinations (CEEB); *Choice Behavior

ABSTRACT

This study assessed the ability of high school advanced placement history students to choose the essay topic on which they can get the highest score. A second, equally important, question was whether the score on the chosen topic was more highly related to other indicators of proficiency in history than the score on the unchosen topic. Overall, for both U.S. and European history, scores were about one-third of a standard deviation higher for the preferred topic than for the other topic. For U.S. history, about 32% of the students made the wrong choice; that is, 32% got a higher score on the other topic than on the preferred topic. In European history, 29% made the wrong choice. In the U.S. history sample, the preferred essay correlated 0.40 with an external criterion score versus 0.34 for the other essay. In the European history sample, the preferred essay correlated 0.52 with the external criterion compared to 0.44 for the other topic. An appendix presents the U.S. and European history standard topics. (Contains four figures, five tables, and five references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-96-4

ED 400 332

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

CHOICE AMONG ESSAY TOPICS: IMPACT ON PERFORMANCE AND VALIDITY

Brent Bridgeman
Rick Morgan
Ming-mei Wang

Educational Testing Service
Princeton, New Jersey
February 1996

BEST COPY AVAILABLE

TM025137.

Choice Among Essay Topics: Impact on Performance and Validity

Brent Bridgeman, Rick Morgan, and Ming-mei Wang

Educational Testing Service

Running Head: CHOICE AMONG ESSAY TOPICS

Copyright © 1996. Educational Testing Service. All rights reserved.

Abstract

This study assessed the ability of history students to choose the essay topic on which they can get the highest score. A second, equally important, question was whether the score on the chosen topic was more highly related to other indicators of proficiency in history than the score on the unchosen topic. Overall, for both US and European history, scores were about one-third of a standard deviation higher for the preferred topic than for the other topic. For US history, about 32% of the students made the wrong choice, that is 32% got a higher score on the other topic than on the preferred topic. In European history, 29% made the wrong choice. In the US history sample, the preferred essay correlated .40 with an external criterion score vs .34 for the other essay; in the European history sample, the preferred essay correlated .52 with the external criterion compared to .44 for the other topic.

Choice Among Essay Topics:
Impact on Performance and Validity

On a number of national examinations, examinees are asked to select one of several possible topics and write an essay on the topic that they have chosen. With different students writing on different topics (essentially taking different tests) it may be difficult or impossible to equate scores across topics (Wainer & Thissen, 1994). If knowledge of a specific narrow subject area or topic is being assessed, it makes little sense to allow choice.

Despite the problems created by allowing choice, not allowing choice may not be a reasonable alternative. If the primary purpose of the assessment is to allow students to demonstrate their ability to organize evidence and present a cogent argument on a subject that they know well, choice may be required for a valid assessment. For example, imagine trying to devise a fair assessment in European history when students in some classes explored the 17th century aristocracy in depth and lightly skimmed over the Lutheran Reformation while other classes had the opposite pattern. An examination with no choice (everybody must write an essay on the Lutheran Reformation) may superficially appear to be fair but would in fact be biased in favor of those students in the classes that gave particular attention to that topic. One possibile alternative to question choice is to include questions only on topics that are known to be covered about equally in all courses. But, as Wainer and Thissen (1994) note, this might discourage teachers from covering topics that are outside of the central core. All courses might be reduced to a limited set of lowest-common-denominator topics. If consequential validity is taken seriously (Messick, 1989), discouraging teachers from pursuing the most educationally beneficial course is a major mark against the validity of an assessment program.

Aside from benefits to the instructional system, topic choice is problematic in a high stakes assessment context if some choices yield higher scores than other choices. The Advanced Placement (AP) Program provides such a context. In this program, colleges extend credit and/or placement into more advanced courses based on a standardized examination given near the end of college-level courses offered in high schools. The examinations typically have a multiple-choice section and a free-response (or essay) section that may allow some choice among topics. In a study of the five AP examinations that allowed choice, Pomplun, Morgan, and Nellikunnel (1992) showed that the average scores on topics chosen by different students varied considerably even when level of performance on the mandatory parts of the examination was held constant. A subsequent study (Morgan, Pomplun, & Nellikunnel, 1993) found that the earlier results were consistent across ethnic and gender subgroups; the various subgroups were equally good (or equally bad) at selecting the topics that yielded the highest scores. However, the wisdom of a particular topic choice for a particular individual could not be evaluated. Even if average scores are generally low on a particular topic, there may still be individuals who are especially familiar with that topic who would get higher scores on that topic than on any of the alternative topics. Wang, Wainer, and Thissen (1993), using a design in which students could express a preference between two multiple-choice questions but were required to answer both of them, found that students frequently made the wrong choice; they got the chosen question wrong and the unchosen question correct. However, this result was based on a multiple-choice examination in which examinees may have been drawn to answer choices that were constructed to appear quite plausible to the partially informed student. Such questions might appear to be easy to poorly prepared students. This is quite unlike the typical choice situation in an essay examination in

which there is no intent to draw students toward questions that appear to be easy but in fact are not. Good data on whether individuals tend to make the correct choice on essay examinations appears to be nonexistent. The current experiment explored the question of whether AP history students who are given a choice of topics tend to make the right choice; that is, do they choose the topic on which they can get the highest score.

If score choice is indeed a valid method for allowing students to more accurately demonstrate their general proficiency in dealing with historical issues, then the score on the chosen topic should be a better indicator of proficiency. Thus, a second, equally important, question for the current research was whether the score on the chosen topic was more highly related to other indicators of proficiency in history than the score on the unchosen topic.

Method

Sample

A random sample of high school teachers with 20 or more students in their college-level Advanced Placement (AP) history courses in United States history or European history were asked to administer specially constructed essay tests in their classrooms within two weeks of the national AP administration in May. Scores on these experimental essays were then matched with scores from the national administration. After matching, the final sample consisted of 538 U.S. history students and 377 European history students.

Materials and procedures

For both subject areas, four essay topics were selected representing different eras and different emphases (e.g., social/intellectual history or political/economic history). The four topics were arranged in four pairs such that each student would have a choice from two topics. Order was counterbalanced so that the topic that was listed first in half of the pairs was second in the other half, although with four topics and four pairings not all possible combinations could be represented. Specifically, Topic 1 was administered together with Topic 3 or Topic 4, and Topic 2 was administered with Topic 3 or Topic 4, but Topic 1 was not administered with Topic 2 and Topic 3 was not administered with Topic 4. (Topics are presented in the Appendix.) Students were told that they should first choose their preferred topic, although they should answer both and both would be scored. In part, the instructions stated, "Read both questions, and choose the question that you are best prepared to answer thoroughly in the time permitted. Circle the number of this question....You should answer both questions, allowing about thirty minutes for each answer." The testing session lasted one hour.

Each essay was holistically scored on a nine-point scale by the same pool of readers, and at the same time, as the regular national AP scoring. Each essay in a pair was read by a different reader, and readers did not know which topic the examinee had selected as the preferred topic.

Scores from the national test. Three scores from the national AP examination were used: the formula score on the 100 multiple-choice questions, the score on the document-based question (DBQ), and the score on the standard essay. A fixed time limit for writing both essays was enforced, but how much time to spend on each essay was merely suggested. For the DBQ, the students had a 15 minute period to study 10 to 20 documents (including paragraphs from

original sources, maps, graphs, drawings, and political cartoons) and then had a suggested time of 50 minutes (in the U.S. History examination) or 45 minutes (in the European History examination) to write an essay showing their ability to formulate an argument and support it with the documentary evidence. (For the U.S. History examination, students were also expected to bring outside knowledge into their answers, but outside knowledge of specific facts was not needed for the European DBQ.) For the standard essay topics (of the type provided in the Appendix), students selected one of 5 (U.S.) or 6 (European) questions presented and had a suggested time of 50 minutes (U.S.) or 45 minutes (European) to compose an essay. The DBQ and standard essays were both evaluated on 15 point scales; the somewhat shorter essays in the experimental administration were evaluated on 9 point scales. A composite score was formed by multiplying the multiple-choice formula score by .9, multiplying the combined essay score (maximum of 15 points on each essay for a total maximum of 30) by 3, and summing the two weighted scores so that the essays and multiple-choice questions each contributed a maximum of 90 points to the composite score.

Results and Discussion

The means and standard for the U.S. History scores from the national administration and the experimental administration are presented in Table 1. For comparison, scores from all 114,475 students who took the national examination are also included. The experimental sample scored slightly higher than the national averages and standard deviations were comparable; the sample appeared to be well within bounds for making meaningful generalizations. Table 1 shows that average scores on the preferred topic were about half a point (0.3 in SD units) higher than scores on the other topic (matched sample $t[537] = 6.0$, $p < .001$). Some of this effect might be

BEST COPY AVAILABLE

attributable to students working harder on their preferred topic. On the other hand, they might also work harder on the other topic to compensate for their perceived weakness.

Table 2 presents comparable information for the European History sample. The sample was very similar to the entire population that took the examination at the national administration. Once again, scores were significantly higher (by about 0.5 SD) on the preferred topic ($t[376] = 8.8$, $p < .001$).

Even if average differences were relatively small, large positive differences (i.e., substantially higher scores on the preferred essay) could still be considerably more common than large negative differences. Figure 1 shows the distribution of the difference scores (score on preferred topic minus score on other topic) for the U.S. History sample. Although the modal difference was 0 (no difference) and there were about as many +1 scores as -1 scores, the more extreme differences clearly favored the preferred topic. For every difference category over 1 in absolute value, there were substantially more people in the plus categories, indicating a higher score on the preferred topic. For example, there were almost twice as many people in the +4 category as in the -4 category. For U.S. history, 52% of the students scored higher on their preferred topic and 32% scored higher on the other topic with the remaining 16% receiving the same score on both topics.

Figure 2 presents comparable information for the European History sample. Differences were even more obviously skewed in the direction of higher scores on the preferred essay. The mode was +1, and there were more than twice as many +2s as -2s. About 58% of the students got a higher score on their preferred topic; only 30% got a higher score on their non-preferred topic.

The U. S. History sample was divided into thirds based on the composite score distribution. Differences between scores on the preferred and other topic were computed separately for men and women within each of these thirds. Box and whisker plots of the difference scores are presented in Figure 3. The lower boundry of each box is the 25th percentile and the upper boundry is the 75th percentile (calculated as Tukey's hinges), and the horizontal line within the box is the median. As the figure shows, differences were comparable for men and women and for students at different history competence levels as indexed by their composite scores. A composite score level by gender ANOVA on the difference scores yielded non-significant results (Fs of less than 1) for both the main effects and the interaction.

Figure 4 is the European History version of Figure 3. Once again gender differences appeared to be minor, but difference scores were not uniform across the different levels of the composite score distribution with slightly larger differences for students with the highest composite scores, as indexed by the higher values for the 25th and 75th percentiles in the top third of the composite score distribution. This was confirmed by a significant effect for composite score level in the ANOVA ($F[2, 371] = 5.3$, $p < .01$), but non-significant results (Fs less than 1) for the gender effect and interaction. The median, however, was insensitive to these differences and remained at 1.0 in each composite score third.

Although the data suggest a general trend for a majority of the students to score higher on their preferred question, the consideration of a particular choice more clearly illustrates the question-level effects. One form had Topic 4 (the Puritan dream in New England) listed first and Topic 2 (the groundwork for the Civil War) listed second. (Topic numbers are provided for reference to the topics as listed in the Appendix; on the student's question booklet they were

called 1 and 2.) As can be seen in the row totals in Table 3, about twice as many people preferred the Civil War topic over the Puritan dream topic, and (from the column total) overall more people received a higher score on the Civil War topic. But, of the students preferring the Puritan dream topic, a few more (22 vs 19) scored higher on that topic than on the Civil War topic. The % wrong column indicates the percent of students who made the wrong choice, that is, who got a higher score on the other topic than on their preferred topic. Over all eight pair-wise topic comparisons (four among U.S. history topics and four among European history topics), the total percent wrong ranged from 19% to 36%. These estimates are conservative (i.e., too high) because some of the apparently wrong choices reflect nothing more than measurement errors caused by fluctuations among readers. Suppose a student wrote equally good essays on both topics, but the reader of the second topic used an unusually strict grading standard and assigned it a lower score. If the student preferred the first topic, this would be counted as a wrong choice, even though with a different rater it might be a correct choice. If we focused on correct choices rather than wrong choices, we would have the same problem in reverse.

Table 4 shows a comparable result for one of the European History forms. The positive effects of choice are particularly evident in this table. Overall, the first topic (Lutheran Reformation) could be considered easier because more students got a higher score on it than on the second topic (English and French aristocracy). Although relatively few students preferred the aristocracy topic over the Lutheran Reformation topic, nearly all students who made that choice clearly were making the correct choice for them. Only 2 of the 24 students who chose the aristocracy topic got a higher score on the Lutheran Reformation topic.

Correlational results

Table 5 shows the correlation of the scores from the preferred and other essay with scores from the national administration. The table also shows the correlation from the higher score on the two essays and from the combination of both scores. Correlations with the composite score are of primary interest because that score is the most reliable indicator of competence over all of the skills assessed in the examination. It is not surprising that the combination of both essays yields the highest correlations, although the nearly equal correlation from the higher of the two essay scores is of interest. If optimum performance is known, adding a score based on sub-optimum performance may be of little benefit. Of course, the problem is that the higher of the two scores cannot be known unless both are administered and scored, resulting in no savings over using the simple sum. However, preference can be known in advance of writing and scoring. The table also suggests that the score on the preferred essay is a better indicator of overall historical competence than is the score on the essay that was not preferred. Although the difference between these correlations falls short of conventional statistical significance in the two samples considered separately, when they are combined (averaged after an r to Z transformation) the difference is significant ($t[909] = 2.0$, $p < .05$).

Conclusion

Allowing choice among essay topics complicates scoring and equating. The potential biases that can result from allowing choice have been well documented (Pomplun, Morgan, & Nellikunnel, 1992; Wainer & Thissen, 1994). Nevertheless, failing to permit choice may disadvantage students who are unfamiliar or uncomfortable with the common topic. In the example above, if the test designer forced all examinees to write on the Lutheran Reformation,

students who could write a better essay on 17th century aristocracy would be disadvantaged. Note that, given limited testing time and scoring resources, someone must choose the essay topic; the question is whether the test designer or examinee should get to make the choice. If the test is intended to assess specific knowledge, such as the details of the Reformation, test designer choice is clearly appropriate. But if the test is designed to assess skill in organizing and presenting historical arguments, a strong case for examinee choice can be made. Not only is the score higher on the chosen essay, but the higher score appears to be a better indicator of competence, at least as indexed by overlap with other scores in the same general domain.

If choice is to be permitted, special efforts are required to make certain that the scoring is as comparable as possible across topics. Techniques for equating scores generated by different topics are not totally satisfactory because they must assume that the other parts of the examination provide all the information that is needed to equate essays. This assumption is clearly not met if the essays are assessing skills that are intentionally different from the skills assessed by the multiple-choice questions. Any equating method should work best in combination with efforts to keep topic difficulty factors from emerging in the first place. Specifically, scoring rubrics should be developed that are applicable to all topics. The precise evidence that needs to be reviewed in a particular answer may vary from topic to topic, but the quantity and quality of required evidence for a particular score should be consistent. This can not be accomplished by separate groups of raters setting different standards for different topics. There must be a single group of raters that establishes a single standard. Even though raters may feel more comfortable specializing in a particular topic, consideration should be given to having all raters read all topics. This would minimize the chances that the readers of a particular topic will inadvertently start

using criteria that are more or less lenient than the criteria used by raters of other topics. If such techniques are adopted, the benefits of score choice can be maximized and the disadvantages minimized.

References

Messick, S. (1989). Validity. In R. L. Linn (Ed.). Educational Measurement (3rd ed., pp. 13-103). New York: Macmillan.

Morgan, R., Pomplun, M., & Nellikunnel, A. (1993). Choice in Advanced Placement tests and subgroup equity (ETS SR-93-167). Princeton, NJ: Educational Testing Service.

Pomplun, M., Morgan, R., & Nellikunnel, A. (1992). Choice in Advanced Placement tests (ETS SR-92-51). Princeton, NJ: Educational Testing Service.

Wainer, H., & Thissen, D. (1994). On examinee choice in educational testing. Review of Educational Research, 64, 159-195.

Wang, X. B., Wainer, H., & Thissen, D. (1993). On the viability of some untestable assumptions in equating exams that allow examinee choice (ETS Tech. Rep. No. 93-31). Princeton, NJ: Educational Testing Service.

Appendix

U. S. History Standard Essay Topics

Topic 1. Assess the impact of THREE of the following on the status of African Americans from the end of Reconstruction to 1900.

The Fourteenth Amendment

"Black Codes"

Plessy v Ferguson

The Atlantic Compromise

Topic 2. Identify THREE of the following and evaluate the relative importance of each of the THREE in laying the groundwork for the Civil War.

Abolitionism

The Mexican War

The Kansas-Nebraska Act

The Dred Scott decision

Topic 3. Analyze the ways in which THREE of the following called into question United States preeminence as a global power.

The postwar reconstruction of Germany and Japan

Nuclear proliferation

The Vietnam War

The Organization of Petroleum Exporting Countries (OPEC)

Topic 4. Analyze the relative importance of religious dissent and demographic change in undermining the Puritan dream of establishing a godly and orderly society in seventeenth century New England.

European History Standard Essay Topics

Topic 1. "In seventeenth-century England the aristocracy lost its privileges but retained its power; in seventeenth-century France the aristocracy retained its privileges but lost its power."

Assess the accuracy of this statement with respect to political events and social developments in the two countries in the seventeenth century.

Topic 2. To what extent and in what ways has twentieth-century physics challenged the Newtonian view of the universe and society?

Topic 3. Assess the strengths and weaknesses of the economic revival of Western Europe between 1945 and 1970.

Topic 4. What were the responses to the Catholic authorities in the sixteenth century to the challenges posed by the Lutheran Reformation?

Copyright © 1993 by Educational Testing Service. All rights reserved. Reproduced by permission.

Author Notes

Thanks to Hunter Breland, Samuel Livingston, and Robert Mislevy for their insightful and useful comments on an earlier draft of this paper. The financial support of the College Board is gratefully acknowledged; the points of view expressed are solely the authors' and do not necessarily represent official College Board position or policy.

Table 1

Means and Standard Deviations for U. S. History Sample

	Population (n=114,475)		Sample (n=538)	
Score	*M*	*SD*	*M*	*SD*
Multiple-choice	49	16	52	15
DBQ	6.1	2.4	6.7	2.3
Std. essay	6.1	2.7	6.7	2.6
Preferred Topic	--	--	4.5	2.1
Other Topic	--	--	3.9	2.1

Table 2

Means and Standard Deviations for European History Sample

	Population (n=30,493)		Sample (n=377)	
Score	*M*	*SD*	*M*	*SD*
Multiple-choice	49	17	47	17
DBQ	7.3	2.4	7.4	2.2
Std. essay	6.7	2.5	6.9	2.6
Preferred Topic	--	--	3.6	1.8
Other Topic	--	--	2.7	1.6

Table 3

Topic Preference by Higher Score for U.S. History Students
Who Wrote on Civil War and Puritan Dream Topics

Preferred Topic	Higher Score on Civil War	Same Score on Both	Higher Score on Puritan dream	Total	% Wrong
Puritan dream	19	8	22	49 (34%)	39
Civil War	48	21	28	97 (66%)	29
Total	67 (46%)	29 (20%)	50 (34%)	146	32

Table 4

Topic Preference by Higher Score for European History Students
Who Wrote on Reformation and Aristocracy Topics

Preferred Topic	Higher Score on Aristocracy	Same Score on Both	Higher Score on Reformation	Total	% Wrong
Reformation	17	13	48	78 (77%)	22
Aristocracy	18	4	2	24 (24%)	8
Total	35 (34%)	17 (17%)	50 (49%)	102	19

Table 5

Correlation of Essay Scores from Experimental Administration with Scores from National Administration

		Essay Scores from Experimental Administration			
Sample	National Scores	Preferred	Other	Higher	Both
U.S.	DBQ	.23	.21	.25	.28
	Std. essay	.23	.22	.29	.29
	Multiple-choice	.39	.32	.45	.46
	Composite	.40	.34	.46	.47
European	DBQ	.28	.30	.31	.34
	Std. essay	.38	.32	.41	.42
	Multiple-choice	.51	.40	.52	.55
	Composite	.52	.44	.54	.57

Note.--*N* is 538 for U.S. History and 377 for European History

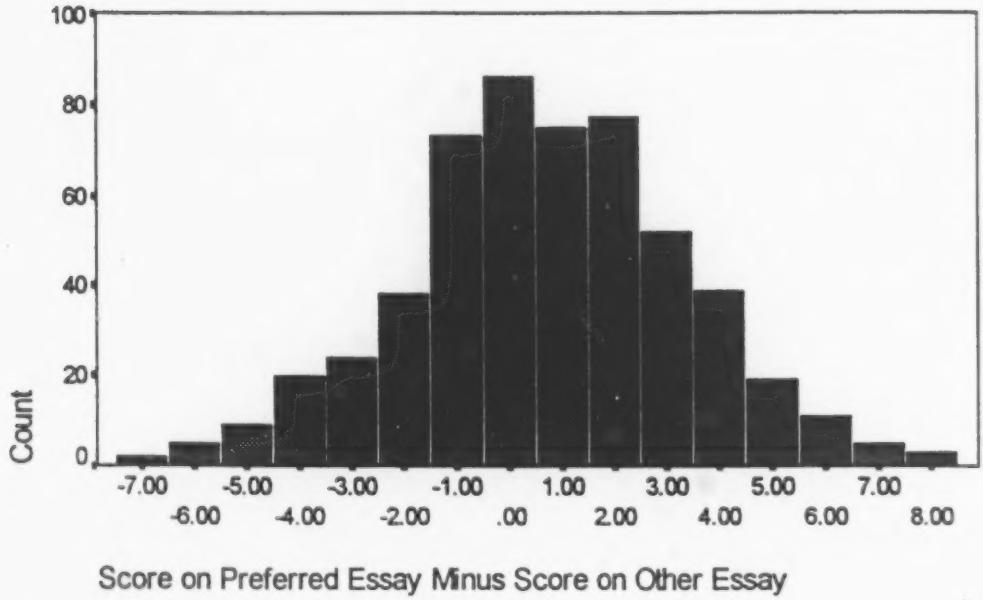

FIGURE 1. U. S. History: Difference between score on preferred essay and score on other essay

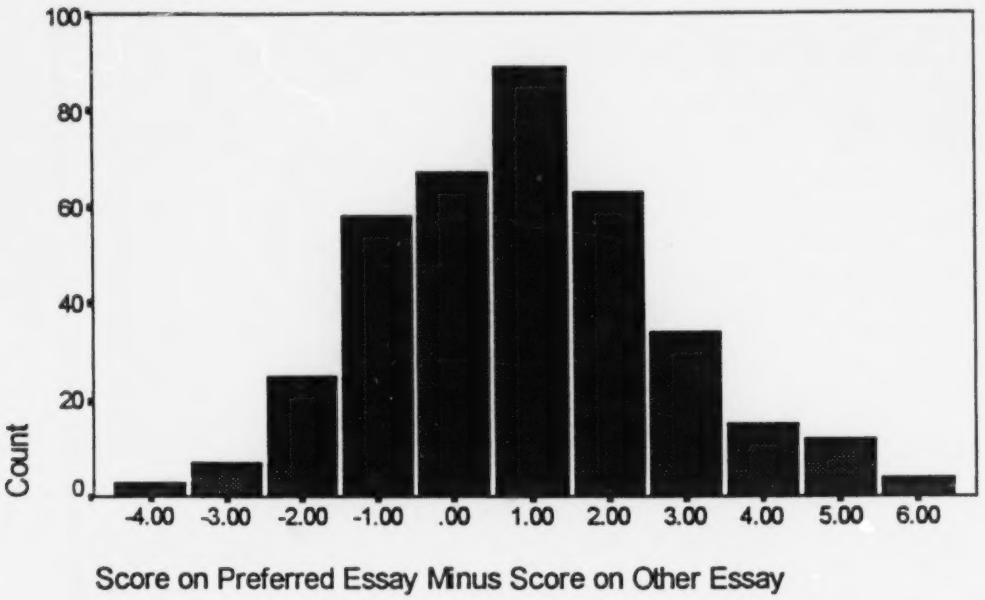

FIGURE 2. European History: Difference between score on preferred essay and score on other essay

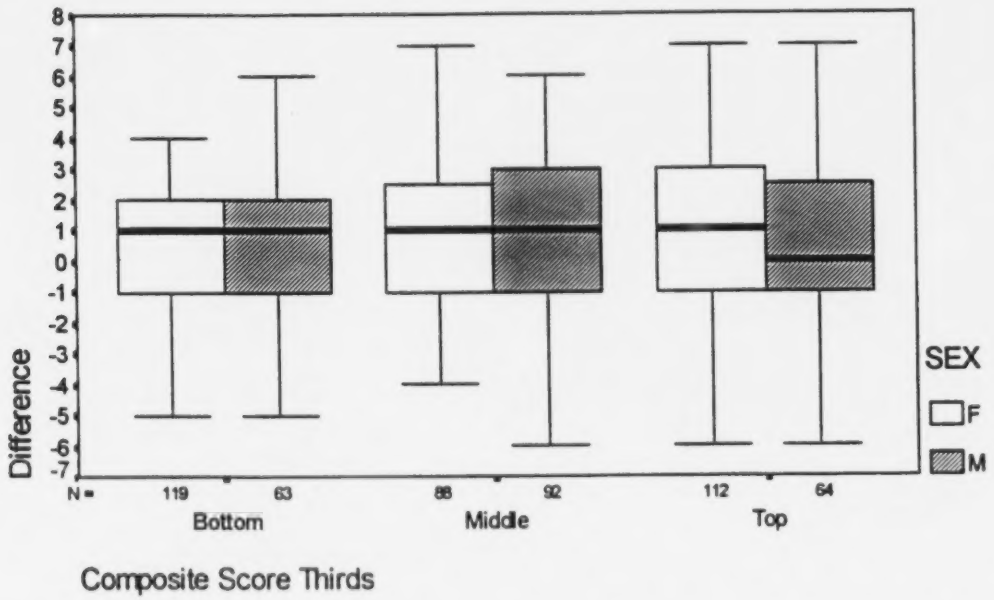

FIGURE 3. U. S. History: Box and whisker plot of difference between score on preferred essay and score on other essay by composite score thirds

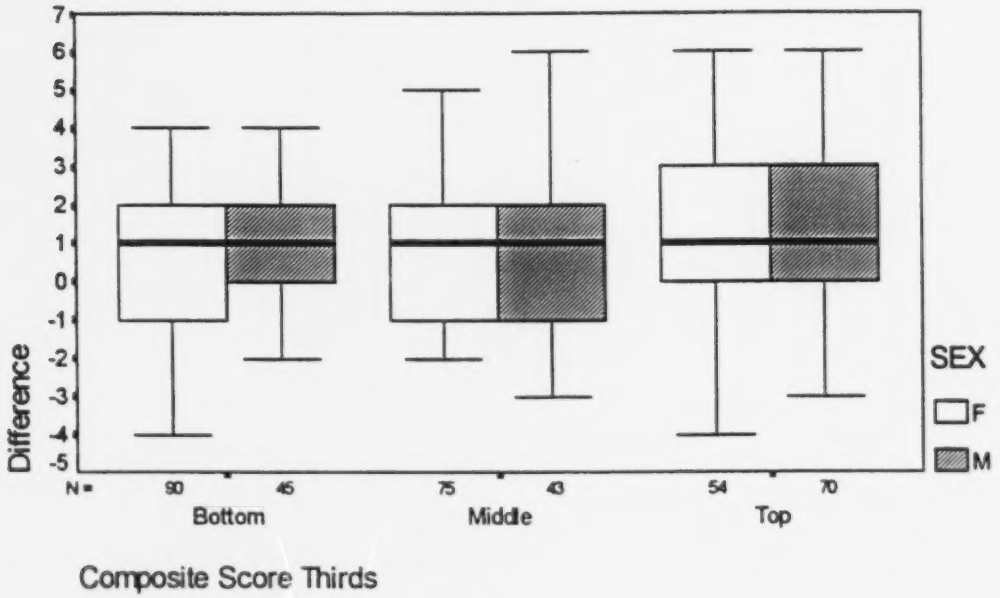

FIGURE 4. European History: Box and whisker plot of difference between score on preferred essay and score on other essay by composite score thirds